Stochastic ranking process and web ranking numbers

Tetsuya Hattori (Keio University)[1]

1 Introduction.

This is a summary of a series of our studies [13, 14, 15, 16, 12] on stochastic ranking process, with its applications on ranking numbers found on the web, such as sales ranks at an online bookstore amazon.co.jp, and thread title listings of an online collected bulletin board 2ch.net. This is a joint work with K. Hattori, Y. Hariya, Y. Nagahata, Y. Takeshima, and T. Kobayashi.

In Section 2, we consider mathematical aspects of stochastic ranking process. We define the stochastic ranking process with the jump times of the particles determined by Poisson random measures, and state that the joint empirical distribution of scaled position and intensity measure converges almost surely in the infinite particle limit. We give an explicit formula for the limit distribution, which can be characterized as a unique global classical solution to an initial value problem for the inviscid Burgers system of non-linear partial differential equations with time dependent coefficients and with evaporation terms. This characterization is in accord with the hydrodynamic limit theories, where a macroscopic time development of collective microscopic random motion of particles is smooth, so that it satisfies a system of partial differential equations.

In Section 3, we show ranking data collected from actual websites at the Amazon online bookstore and at an online collected bulletin board 2ch.net, and show how they are explained by the properties of stochastic ranking process given in Section 2. It is a new social phenomena to have a large number of items aligned dynamically in an order of popularity, and real time values of ranks of thousands can be observed. By performing a statistical fit of the data to the formulas from the stochastic ranking process, one can analyze a 'long tail' structure of social activities at these websites. We conclude that the best hit or top sales items dominate the activities both at Amazon.co.jp and 2ch.net, so that, in particular, Amazon.co.jp, perhaps in contrast to its fame, is not an example of a long tail business.

2 Stochastic ranking process.

The latest version of stochastic ranking process, which extends the original model [13] to the case of time dependent intensities, is defined as follows [12]. Let $\mathcal{M}(\mathbb{R}_+)$ be the space of Radon measures ρ on the Borel σ-algebra $\mathcal{B}(\mathbb{R}_+)$ of non-negative reals $\mathbb{R}_+$. Let N be a positive integer, and let $\nu_i^{(N)}$, $i = 1, 2, \cdots, N$, be independent Poisson random measures (Poisson point processes) on $\mathbb{R}_+$, defined on a probability space $(\mathrm{P}, \mathcal{F}, \Omega)$. For each i, denote the intensity measure of $\nu_i^{(N)}$ by $\rho_i^{(N)}$;

$$\mathrm{E}[\ \nu_i^{(N)}(A)\ ] = \rho_i^{(N)}(A),\ A \in \mathcal{B}(\mathbb{R}_+). \tag{1}$$

We assume that $\rho_i^{(N)} \in \mathcal{M}(\mathbb{R}_+)$ and that $\rho_i^{(N)}$ is continuous (i.e., $\rho_i^{(N)}(\{t\}) = 0$, $t \geqq 0$) for all N and i. Let $x_1^{(N)}, x_2^{(N)}, \cdots, x_N^{(N)}$ be a permutation of $1, 2, \cdots, N$, and

[1] More materials available at `http://web.econ.keio.ac.jp/staff/hattori/amazone.htm`

define a process $X^{(N)} = (X_1^{(N)}, \cdots, X_N^{(N)})$ by

$$\begin{aligned} & X_i^{(N)}(t) \\ & = x_i^{(N)} + \sum_{k=1}^{N} \int_0^t \mathbf{1}_{X_k^{(N)}(s-0) > X_i^{(N)}(s-0)} \, \nu_k^{(N)}(ds) + \int_0^t (1 - X_i^{(N)}(s-0)) \, \nu_i^{(N)}(ds), \\ & i = 1, 2, \cdots, N, \ t \geqq 0 \, , \end{aligned} \tag{2}$$

where, $\mathbf{1}_A$ is the indicator function of an event A. We call the process $X^{(N)}$ defined by (2), a stochastic ranking process, after [13, 14, 15].

Denote the unit measure concentrated on c by δ_c. With probability 1 we can write

$$\nu_i^{(N)} = \sum_{j=1}^{\infty} \delta_{\tau_{i,j}^{(N)}}, \ i = 1, 2, \cdots, N, \tag{3}$$

where, with probability 1, $\tau_{i,j}^{(N)}$'s are random variables satisfying $0 < \tau_{i,1}^{(N)} < \tau_{i,2}^{(N)} < \cdots$, $i = 1, 2, \cdots, N$, and $\tau_{i,j}^{(N)} \neq \tau_{i',j'}^{(N)}$ if $(i,j) \neq (i',j')$. In the following, we work on the event that these inequalities hold. $X_i^{(N)}(t)$ has an explicit expression using $\tau_{i,j}^{(N)}$'s:

$$X_i^{(N)}(t) = \begin{cases} x_i^{(N)} + \displaystyle\sum_{i';\, x_{i'}^{(N)} > x_i^{(N)}} \mathbf{1}_{\tau_{i',1}^{(N)} \leqq t} & 0 \leqq t < \tau_{i,1}^{(N)} \, , \\ 1 + \displaystyle\sum_{i'=1}^{N} \mathbf{1}_{\exists j' \in \mathbb{N};\, \tau_{i,j}^{(N)} < \tau_{i',j'}^{(N)} \leqq t} & \tau_{i,j}^{(N)} \leqq t < \tau_{i,j+1}^{(N)}, \ j = 1, 2, 3, \cdots, \end{cases} \tag{4}$$

for $i = 1, \cdots, N$.

In the time homogeneous case, namely, the case where there exists positive constants $w_i^{(N)}$ such that $\rho_i^{(N)}((0,t]) = w_i^{(N)} t$ for $t \geqq 0$, a discrete time version of the process (4) has been known for a long time [26, 23, 17, 6, 22, 21] and is called move-to-front (MTF) rules. The process has, in particular, been extensively studied as a model of least-recently-used (LRU) caching in the field of information theory [24, 8, 4, 7, 5, 25, 9, 11, 10, 18, 19, 20], and also is noted as a time-reversed process of top-to-random shuffling.

Put

$$X_C^{(N)}(t) = \sum_{i=1}^{N} \mathbf{1}_{\tau_{i,1}^{(N)} \leqq t}, \quad t \geqq 0. \tag{5}$$

$X_C^{(N)}(t)$ is a random variable which denotes the position of the boundary between the top side $x \leqq X_C^{(N)}(t)$ and the tail side $x > X_C^{(N)}(t)$, where each particle in the top side (i.e., i which satisfies $X_i^{(N)}(t) \leqq X_C^{(N)}(t)$) has experienced jump to the top by time t (i.e., $\tau_{i,1}^{(N)} \leqq t$), and the particles in the tail side are those particles which have not jumped to the top by time t.

Proposition 1 ([13, Prop. 2],[12, Prop. 1.1, Cor. 1.2]) *Let $t \geqq 0$. Assume that a sequence of distributions $\{\lambda_t^{(N)} \mid N \in \mathbb{N}\}$ on $\mathbb{R}_+$ defined by*

$$\lambda_t^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\rho_i^{(N)}((0,t])} \tag{6}$$

converges weakly as $N \to \infty$ to a probability distribution λ_t . Then the scaled position of the boundary

$$Y_C^{(N)}(t) = \frac{1}{N} X_C^{(N)}(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\tau_{i,1}^{(N)} \leqq t} \tag{7}$$

converges almost surely as $N \to \infty$ to

$$y_C(t) = 1 - \int_0^\infty e^{-s} \lambda_t(ds). \tag{8}$$

Assume furthermore that λ_t is continuous in t with respect to the topology of weak convergence. Then for almost all sample $\omega \in \Omega$, $Y_C^{(N)}(\cdot, \omega) : \mathbb{R}_+ \to [0,1)$ defined by (7) converges pointwise in t as $N \to \infty$ to a deterministic function $y_C : \mathbb{R}_+ \to [0,1)$ defined by (8). ◇

Consider a joint empirical distribution $\mu^{(N)}$ of intensity measure $\rho_i^{(N)}$ and scaled position

$$Y_i^{(N)}(t) = \frac{1}{N}(X_i^{(N)}(t) - 1), \tag{9}$$

defined by

$$\mu_t^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\rho_i^{(N)}, Y_i^{(N)}(t))}, \quad t \geqq 0. \tag{10}$$

$\mu_t^{(N)}$, $N \in \mathbb{N}$, are random variables whose samples are distributions on the product space $\mathcal{M}(\mathbb{R}_+) \times [0,1)$ of space of Radon measures $\mathcal{M}(\mathbb{R}_+)$ and an interval $[0,1) \subset \mathbb{R}_+$. We consider the standard vague topology on $\mathcal{M}(\mathbb{R}_+)$. Since $\mathbb{R}_+$ is a Polish space, i.e., complete and separable metric space, so is $\mathcal{M}(\mathbb{R}_+)$ [2, Theorem 31.5], and consequently, $\mathcal{M}(\mathbb{R}_+) \times [0,1)$ is also a Polish space [2, Example 26.2].

Assume that a sequence of initial configurations

$$\mu_0^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\rho_i^{(N)}, N^{-1}(x_i^{(N)} - 1))}, \quad N = 1, 2, \cdots,$$

converges weakly as $N \to \infty$ to a probability distribution μ_0 on $\mathcal{M}(\mathbb{R}_+) \times [0,1)$. Then, in particular,

$$\Lambda^{(N)}(d\rho) := \mu_0^{(N)}(d\rho \times [0,1)) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\rho_i^{(N)}}(d\rho) \ \to \ \Lambda(d\rho) := \mu_0(d\rho \times [0,1)),$$
$$\text{weakly, as } N \to \infty. \tag{11}$$

Define, for $0 \leqq s \leqq t$,

$$\lambda_{s,t}^{(N)} = \int_{\mathcal{M}(\mathbb{R}_+)} \delta_{\rho((s,t])} \Lambda^{(N)}(d\rho). \tag{12}$$

Note that $\lambda_t^{(N)} = \lambda_{0,t}^{(N)}$ in (6).

Theorem 2 ([13, Thm. 1.5], [12, Thm. 1.3]) *Assume that $\mu_0^{(N)} \to \mu_0$ weakly as $N \to \infty$ for a probability distribution μ_0 on $\mathcal{M}(\mathbb{R}_+) \times [0,1)$. Assume that for each (s,t) satisfying $t \geqq s \geqq 0$,*

$$\lambda_{s,t}^{(N)} \ \to \ \lambda_{s,t} := \int_{\mathcal{M}(\mathbb{R}_+)} \delta_{\rho((s,t])} \Lambda(d\rho), \quad \textit{weakly as} \ \ N \to \infty, \tag{13}$$

where Λ is as in (11). Then for any $t>0$, and for almost all sample $\omega\in\Omega$, the distribution $\mu_t^{(N)}(\omega)$ converges weakly to a non-random probability distribution μ_t on $\mathcal{M}(\mathbb{R}_+)\times[0,1)$.

μ_t has a following expression in terms of $U(d\rho,y,t):=\mu_t(d\rho\times[y,1))$.

$$U(d\rho,y,t):=\mu_t(d\rho\times[y,1))=\begin{cases} e^{-\rho((t-t_0(y,t),t])}\,\Lambda(d\rho) & 0\leqq y\leqq y_C(t),\\ e^{-\rho((0,t])}\,U(d\rho,\hat{y}(y,t),0) & y_C(t)\leqq y<1.\end{cases}\tag{14}$$

Here, $t_0(y,t)$ is the inverse function with respect to t_0 of

$$y_A(t_0,t)=1-\int_{\mathcal{M}(\mathbb{R}_+)} e^{-\rho((t-t_0,t])}\,\Lambda(d\rho),\quad 0\leqq t_0\leqq t,\tag{15}$$

and $\hat{y}(y,t)$ is the inverse function with respect to y of

$$y_B(y,t)=1-\int_{\mathcal{M}(\mathbb{R}_+)} e^{-\rho((0,t])}\,\mu_0(d\rho\times[y,1)),\quad t\geqq 0,\ 0\leqq y<1.\tag{16}$$

◇

Note that $y_C(t)=y_A(t,t)=y_B(0,t)$.

If we impose additional conditions, we may go further for Theorem 2 and prove almost sure convergence as a sequences of processes $\mu^{(N)}\to\mu$, on a finite time interval $[0,T]$. See [12, §4].

The structure of the explicit limit formula (14), in particular, the appearance of the inverse functions t_0 of y_A and $\hat{y}$ of y_B , can mathematically be understood through a system of partial differential equations. Consider the case that the limit distribution Λ is supported on a discrete set: $\Lambda=\displaystyle\sum_\alpha r_\alpha\delta_{\rho_\alpha}$. Then (14) implies, for $U_\alpha(y,t):=\mu_t(\{\rho_\alpha\}\times[y,1))$,

$$U_\alpha(y,t)=\begin{cases} r_\alpha\,e^{-\rho_\alpha((t-t_0(y,t),t])} & 0\leqq y\leqq y_C(t),\\ U_\alpha(\hat{y}(y,t),0)\,e^{-\rho((0,t])} & y_C(t)\leqq y<1,\end{cases}\tag{17}$$

where t_0 and $\hat{y}$ are inverse functions, respectively, of $y_A(t_0,t)=1-\displaystyle\sum_\alpha r_\alpha e^{-\rho_\alpha((t-t_0,t])}$, and $y_B(y,t)=1-\displaystyle\sum_\alpha U_\alpha(y,0)e^{-\rho_\alpha((0,t])}$.

Theorem 3 ([14, Thm. 1], [12, Thm. 1.4]) *Let k be a positive integer, and for each $\alpha=1,2,\cdots,k$, let r_α be a positive constant, $w_\alpha:\ \mathbb{R}_+\to\mathbb{R}_+$ a measurable function satisfying $w_\alpha(t)>0$, $t\geqq 0$, and $u_\alpha:\ [0,1)\to\mathbb{R}_+$ a non-negative smooth strictly decreasing function, satisfying*

$$\sum_{\beta=1}^{k} r_\beta=1,\quad \sum_{\beta=1}^{k} r_\beta w_\beta(t)<\infty,\ t\geqq 0,\quad \textit{and}\quad \sum_{\beta=1}^{k} u_\beta(y)=1-y,\ 0\leqq y<1.\tag{18}$$

Then an initial value problem for a system of first order non-linear partial differential equations (inviscid Burgers equations with a term representing evaporation)

$$\begin{aligned}&\frac{\partial U_\alpha}{\partial t}(y,t)+\sum_{\beta=1}^{k} w_\beta(t)\,U_\beta(y,t)\,\frac{\partial U_\alpha}{\partial y}(y,t)=-w_\alpha(t)U_\alpha(y,t),\\ &(y,t)\in[0,1)\times\mathbb{R}_+,\ \alpha=1,2,\cdots,k,\end{aligned}\tag{19}$$

with a boundary condition

$$U_\alpha(0,t) = r_\alpha,\ t \geqq 0,\ \alpha = 1,2,\cdots,k, \tag{20}$$

and initial data

$$U_\alpha(\cdot,0) = u_\alpha,\ \alpha = 1,2,\cdots,k, \tag{21}$$

has a unique time global classical solution, whose formula is given by (17) with

$$\rho_\alpha((s,t]) = \int_s^t w_\alpha(u)\,du \quad \textit{and} \quad U_\alpha(y,0) = u_\alpha(y). \tag{22}$$

$\diamond$

The system (19) of partial differential equations is solved by a method of characteristic curves, and y_A, y_B, and y_C turn out to be the characteristic curves for (19), which mathematically explains how the inverse functions of these functions appear in the solutions. Theorem 3 indicates that the limit in Theorem 2 has an interpretation that a collective random motion of particles is macroscopically observed as a smooth time development explained by a system of partial differential equations, as in the theory of hydrodynamic limit.

3 Web rankings.

With great advance in the internet technologies, a new application of the process appeared [14, 15, 12]. The mathematical results on the stochastic ranking process have successfully been applied to statistical explanation of practical ranking data, such as the ranking numbers of books found in the web pages of an online bookstore Amazon.co.jp [15, 14], or the order of the subject titles in the title listing pages of a collected web bulletin board 2ch.net [14, 12]. A ranking of a book at Amazon.co.jp jumps close to top of the ranking whenever the book is sold at Amazon.co.jp [15], and a subject title in the web page for the list of 2ch.net jumps to the top whenever a comment (a 'response') concerning the subject is submitted [14]. It turned out that the time developments of the ordering of items on these online systems are found to follow the predictions of the model.

One may wonder why such a simple model as introduced in Section 2 could be observed in actual social activities. An explanation is that the ranking numbers on the web (such as those representing the books, in the case of online bookstores) usually seek to align the web pages in the order of *current popularity* of the pages. A social impact of the development of web-based activities is that it has become possible to catalog a huge amount of unpopular items [1]. In fact, a majority of books catalogued on an online bookstore are sold less than one copy a month. For such books, any reasonable order reflecting the current popularity would be equal to the order of the time of most recent sales, because the second recent sale of such book would be long ago, hence would not reflect current popularity. Thus the move-to-front rule will provide a simple but *universal* model in the rankings on the web.

Note that (2) implies the Markov property

$$\begin{aligned} X_i^{(N)}(t+u) = X_i^{(N)}(u) &+ \sum_{k=1}^N \int_0^t \mathbf{1}_{X_k^{(N)}(s+u-0) > X_i^{(N)}(s+u-0)}\ \tilde{\nu}_k^{(N)}(ds) \\ &+ \int_0^t (1 - X_i^{(N)}(s+u-0))\,\tilde{\nu}_i^{(N)}(ds), \end{aligned}$$

where we put $\tilde{\nu}_i^{(N)}(A) = \nu_i^{(N)}(A+u)$. In practical application, this property enables us to shift the time origin $t = 0$ to the time that a particle we observe jumps to the top, namely, we may set $X_i^{(N)}(0) = x_i^{(N)} = 1$, by adjusting the 'clock' for the intensity measure accordingly. (See Fig. 5 and Fig. 7, as well as [14, 15].)

Note also that if $x_i^{(N)} = 1$, then up to the first jump of i to the top, namely, for $t < \tau_{i,1}^{(N)}$, comparison of (4) and (5) leads to

$$X_i^{(N)}(t) = X_C^{(N)}(t) + 1,$$

Therefore, in practical application in Section 3.2, we may proceed with observing a trajectory (time development) of a single particle, putting the time of its first jump to top as $t = 0$ and observing until its next jump to top, and then apply Proposition 1.

Concerning the explicit time dependence of intensity measures for the Poisson random measures, one should note that data from an online bookstore and from a collected web bulletin board arise as results of social activities, which are expected to contain day-night difference in the intensity. In Section 3.1, we summarize a simple method of [12, §A, §5], to factorize the time dependence and the distribution of relative jump rates among different particles. Then we show the data from amazon.co.jp in Section 3.2, and the data from 2ch.net in Section 3.3, together with statistical applications of the theoretical results.

3.1 Intensities with common time dependence.

In practical situation, intensity measures $\rho_i^{(N)}$ are usually unknown quantities to be determined statistically from observed data. This is usually a difficult task if intensity measures have time dependence, because then we have to consider both particle dependence and time dependence at once in the statistical analysis. Explicit

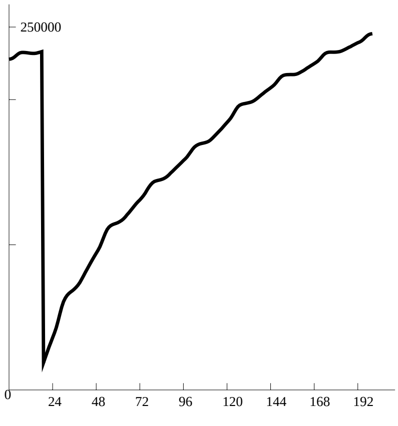

Fig 1:

time dependence, reflecting day-night difference of social activities, are observed in actual data. Fig. 1 is an 8 days plot of Amazon.co.jp rankings for a book. (See [15] for basic fact about Amazon.co.jp ranking and its relation to the stochastic ranking process.) The vertical axis stands for the ranking number, and the horizontal axis is the time axis labelled in the unit of hour. The large discontinuous drops near to the top ranking correspond to the point of sales of the book. Note the 24 hours periodic time dependence. Without explicit time dependence of activities (i.e., the homogeneous case), (28) implies that the curve in Fig. 1 should be concave,

$$y_C''(t) < 0. \tag{23}$$

However, the curve in Fig. 1 actually has convex intervals every 24 hours, proving explicit time dependence, which naturally can be interpreted as day-night difference in social activities.

A simplest way to take day-night-difference of social activity into account, is to assume a common time dependence. Assume that there exist $\tilde{a} \in L^1_{loc}(\mathbb{R}_+)$ and positive constants $w_i^{(N)} > 0,\ i = 1, 2, \cdots, N,\ N = 1, 2, \cdots$, such that the intensity measure (1) is given by

$$\rho_i^{(N)}((s,t]) = w_i^{(N)} \int_s^t \tilde{a}(u)\,du, \quad i = 1, 2, \cdots, N,\ N = 1, 2, \cdots. \tag{24}$$

Proposition 4 *Let $\tilde{a} \in L^1_{loc}(\mathbb{R}_+)$. If there exists a probability distribution λ on $\mathbb{R}_+$ such that*

$$\lambda^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{w_i^{(N)}} \to \lambda, \quad \textit{weakly, as} \quad N \to \infty, \tag{25}$$

then Proposition 1 holds with (24), and $y_C(t)$ of (8) is given by

$$y_C(t) = 1 - \int_{\mathbb{R}_+} e^{-w\,A(t)}\,\lambda(dw), \tag{26}$$

where

$$A(t) = \int_0^t \tilde{a}(u)\,du. \tag{27}$$

◇

The formula (26) is to be compared with the case of the (homogeneous) Poisson process in [13, Proposition 2], where we have

$$y_C(t) = 1 - \int_{\mathbb{R}_+} e^{-wt} \lambda(dw). \tag{28}$$

3.2 Factorization of day-night social activity difference, and sales ranks of Amazon.co.jp .

We can show that under the common time dependence assumption (24), periodic time dependence of $\tilde{a}$ can be factorized, and that the use of (28) is justified in obtaining λ statistically from data. Assume that there exists a positive constant T such that $\tilde{a}(t+T) = \tilde{a}(t),\ t \geqq 0$. We may normalize $w_i^{(N)}$'s in (24) so that $\displaystyle\frac{1}{T}\int_0^T \tilde{a}(u)\,du = 1$ holds. Then $A_p(t) := A(t) - t = \displaystyle\int_0^t (\tilde{a}(u) - 1)\,du$ is a periodic function with period T, and (26) is

$$y_C(t) = 1 - \int_{\mathbb{R}_+} e^{-w\,(t + A_p(t))}\,\lambda(dw). \tag{29}$$

If we collect data at each fixed time of the day, at $t_n = t_0 + nT$, $n = 0, 1, 2, \cdots$, then (29) implies

$$y_C(t_n) = 1 - \int_{\mathbb{R}_+} e^{-w(nT+t_0+A_p(t_0))}\,\lambda(dw). \tag{30}$$

Hence the effect of day-night difference in $\tilde{a}$ is absorbed in the translation of origin of time $t_0 \mapsto t_0 + A_p(t_0)$, and the use of formula (28) is justified.

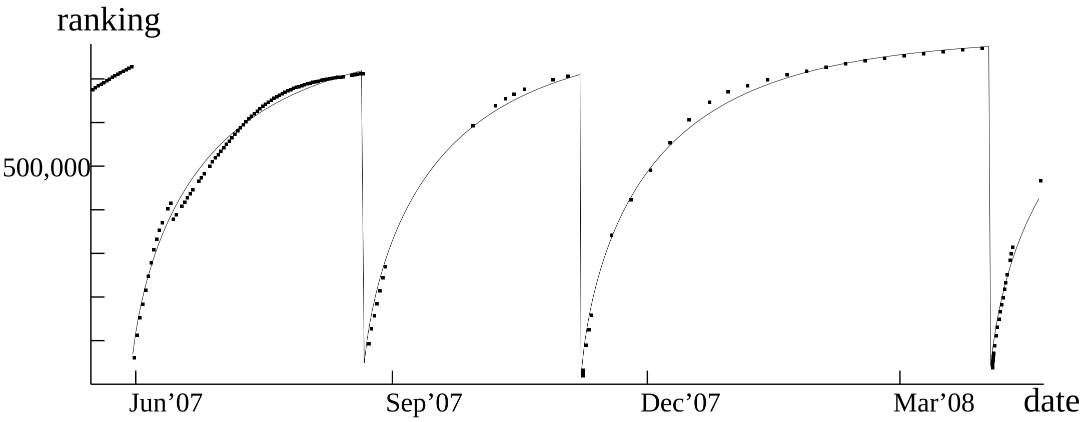

Fig 2:

Fig. 2 is a plot of Amazon.co.jp rankings for a book over a year [15]. The data was taken manually for a year starting in May 2007, at 21:00 each day. As seen in Fig. 2, for a relatively unpopular book, a book which sells less than a copy per week, ranking fall (increase in number) steadily and smoothly at several hundred thousands for much of the time, but once in a while they make sudden jumps to numbers around ten thousand. These occasional large discontinuous jumps near to the top ranking correspond to the point of sales of the book [15].

To apply (28) or (30) to the data, we need to specify λ. A standard choice in social and economic studies seems to be the Zipf’s law, defined by

$$w_i^{(N)} = a\left(\frac{N}{i}\right)^{1/b}, \quad i = 1, 2, \cdots, N, \tag{31}$$

for positive constants a and b. The corresponding $N \to \infty$ weak limit is the (generalized) Pareto distribution, defined by

$$\lambda([w,\infty)) = \begin{cases} \left(\dfrac{a}{w}\right)^b & w \geqq a, \\ 1 & w < a. \end{cases} \tag{32}$$

Substituting (32) in (28), we have

$$y_C(t) = 1 - e^{-at} + (at)^b\Gamma(1-b, at). \tag{33}$$

where Γ is the incomplete Gamma function defined by $\Gamma(z,p) = \displaystyle\int_p^\infty e^{-x}x^{z-1}dx$.

Using the data $\{x_i \mid i = 1, 2, \cdots, n_d\}$ of size $n_d = 77$ at the leftmost arc in Fig. 2, taken between May, 2007 and August, 2007, at 21:00 each day, and choosing to minimize

$$E = E(N, a, b) = \sum_{i=1}^{n_d} \frac{(x_i - N\,y_C(t_i))^2}{x_i}, \tag{34}$$

we obtained the best fit for the parameter set

$$(N^*,\ a^*,\ b^*) = (8.15 \times 10^5,\ 5.30 \times 10^{-4},\ 0.767), \tag{35}$$

with $E_{min} = E(N^*, a^*, b^*) = 4.17 \times 10^4$. In particular, we have $b^* < 1$, which implies that amazon.co.jp earns dominantly from a small number of best hit books [15], rather than the majority of books in the long tail, in contrast to the Amazon bookstores fame as a successful long tail business model [1].

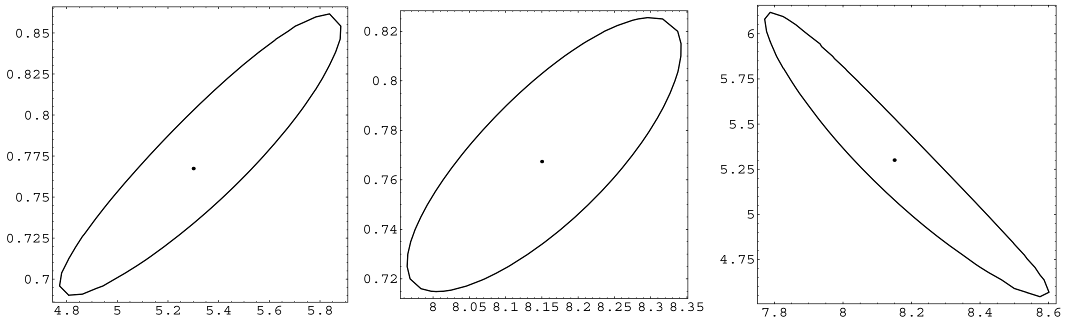

Fig 3:

Fig. 3 shows contour plots of E in (34), representing error estimates (confidence intervals) for the parameters in (35): $E(N, a, b) = \dfrac{\kappa}{n_d}\, E_{min}$, with κ defined (as usual) by $p = \mathrm{P}[\ \chi^2_{n_d} \leqq \kappa\ ]$, where $\chi^2_{n_d}$ is a random variable with chi-square distribution of degree of freedom n_d. The curves in the graphs correspond to the confidence level of 90% , namely, $p = 0.9$. The three figures are cross sections of $N = N^*$, $a = a^*$, $b = b^*$, respectively, in the 3-dimensional parameter space (N, a, b). Horizontal and vertical axes are respectively $a \times 10^4$ and b for the first figure, $N \times 10^{-5}$ and b for the second figure, and $N \times 10^{-5}$ and $a \times 10^4$ for the third figure. The dot in the center of each figure is the best fit (35). Fig. 3 supports $b < 1$, a standard best hit business model, rather than a long tail business model.

To see the stability of the parameters, a similar fit by adding to above mentioned data of size 77 a data of size 21 at the rightmost arc in Fig. 2, taken between November, 2007 and March, 2008, at 21:00 on every Saturday. The best fit is

$$(N^*,\ a^*,\ b^*) = (7.97 \times 10^5,\ 5.93 \times 10^{-4},\ 0.809). \tag{36}$$

The solid curve in Fig. 2 shows the theoretical curve $N\, y_C(t)$ with $y_C(t)$ as in (33) with parameters (36).

We have less data for amazon.com, the original Amazon online bookstore in USA. Fig. 4 shows a data from Amazon.com, which obviously shows a similar behavior as amazon.co.jp data Fig. 2.

3.3 Time change according to intensity measure, and title listings of 2ch.net .

Fig. 5 shows a data from a web page for a list of subject titles at a collected bulletin board 2ch.net. Each curve corresponds to the position of the title of a subject ('thread') in the page of list of titles. See [14] for a description of how the order of

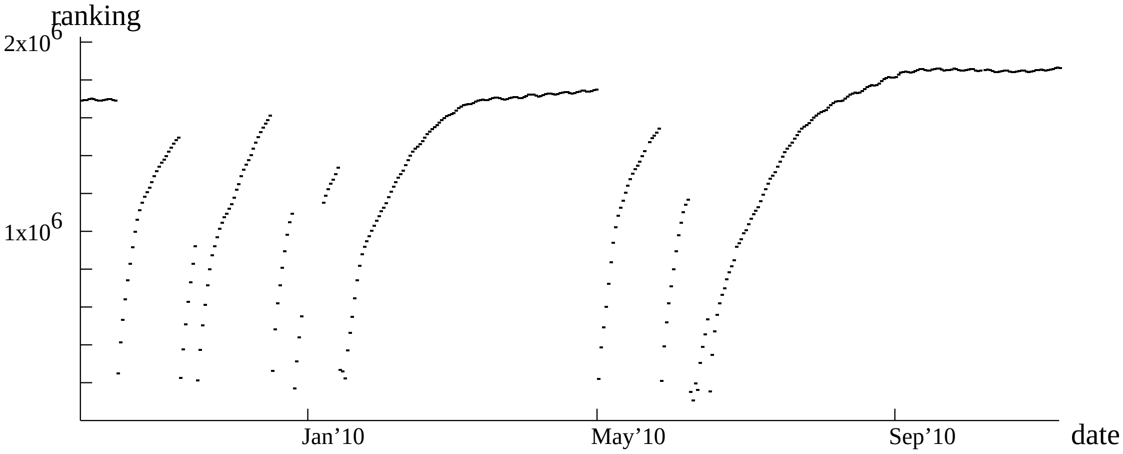

Fig 4:

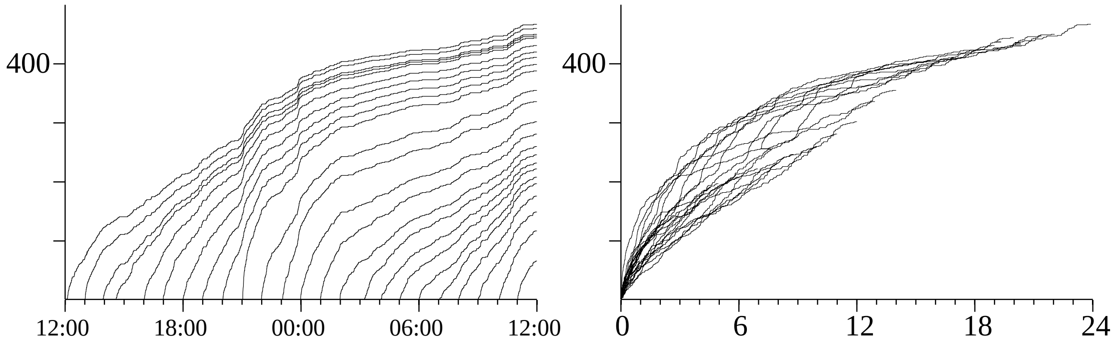

Fig 5:

the list of subject titles is organized at 2ch.net. In short, a thread title jumps to the top of the list if and only if someone writes on the thread (a 'response'), and the jump occurs instantaneously. If one assumes that responses are independent and random, then the time dependence of the thread list is a sample of the stochastic ranking process. The data is taken by Y. Takeshima for 24 hours starting on Oct. 18, 2008, 12:00 JST, using his original data collection program (master thesis, [12]). The vertical axis stands for the position in the list (the horizontal axis at the bottom stands for the top of the list), and the horizontal axis is the time axis labelled in the unit of hour. For clarity of the figure, 24 threads are chosen out of 697 threads, and for each thread, shown is the part from the last jump of the thread until the end of data collection. The second figure is a plot of same data as the first figure, but each curve is shifted in the horizontal direction, so that all the curves starts from the origin $(0, 0)$.

If the jump rate of threads are constant, then the formula (28) for the homogeneous case should be applicable, and the curves, when shifted in the horizontal direction so that the curves start from the origin $(0, 0)$, should follow a single curve defined by (28). As seen from the second figure in Fig. 5, the time dependence of position of threads do not follow a single curve. Also, the first figure in Fig. 5 clearly indicates a violation of convexity (23) at around 21:00 and 00:00. We may interpret the result as internet activities at 2ch.net being more active at night before midnight, compared to deep in the night until early in the morning.

Let us consider the factorization assumption of (24). Since the assumption is neither of logical consequence of the model nor the established social fact, such assumption should be tested by actual application of the formula to the data. Since 2ch.net is very 'transparent', concerning basic facts such as the number of threads in a board or the records of response (jump) times, it is a useful website to test (24), or any other possible practical assumptions.

For $t \geqq 0$, let

$$S^{(N)}(t) = \sum_{i=1}^{N} \nu_i^{(N)}((0,t]) \tag{37}$$

and denote its right continuous inverse by

$$s^{(N)}(t) = \inf\{s \geqq 0 \mid S^{(N)}(s) > t\}. \tag{38}$$

Let $\tilde{a} \in L^1_{loc}(\mathbb{R}_+)$. For simplicity, assume further that

$$\tilde{a}(t) > 0,\ t \geqq 0. \tag{39}$$

Then $A(t)$ of (27) is strictly increasing, and the inverse function A^{-1} is continuous.

Theorem 5 ([12, Thm 5.3, Lem. 5.4]) *Let $\tilde{a} \in L^1_{loc}(\mathbb{R}_+)$, and assume (39). Put*

$$Z(N) = \sum_{i=1}^{N} w_i^{(N)} \tag{40}$$

and assume $\lim\limits_{N\to\infty} Z(N) = \infty$. If, as in Proposition 4, there exists a probability distribution λ on $\mathbb{R}_+$ such that (25) holds, then for each $t \geqq 0$,

$$\frac{1}{Z(N)} S^{(N)}(t) \to A(t), \quad \textit{and} \quad s^{(N)}(Z(N)\,t) \to A^{-1}(t), \quad \textit{in probability, as} \quad N \to \infty, \tag{41}$$

and

$$Y_C^{(N)}(s^{(N)}(Z(N)\,t)) \to y_C(A^{-1}(t)) = 1 - \int_{\mathbb{R}_+} e^{-w\,t}\lambda(dw), \tag{42}$$

in probability, as $N \to \infty$, where $Y_C^{(N)}$ is defined in (7). ◇

Fig. 6 shows the cumulative total number of jumps $S^{(N)}(t)$ in (37) up to time t, for $N = 697$ threads at 2ch.net. The data is from the same board at same time as the data for Fig. 5, collected by Y. Takeshima, and Fig. 6 is accumulated by T. Kobayashi (master thesis, [12]). The dashed line denotes the hypothetical case of constant jump rates. The data is consistent with the observation made for Fig. 5 that the activities (responses) are high at night before midnight, and low between deep in the night to early in the morning.

Fig. 7 is a plot of the same data as Fig. 5, except that the horizontal axis is measured by $S^{(N)}(t)$ of Fig. 6. Fig. 7 is a revised plot of the original one by T. Kobayashi (master thesis, [12]). Compared with the second figure in Fig. 5, the second figure in Fig. 7 is apparently closer to a single curve, which supports an approximate validity of the common time dependence assumption (24).

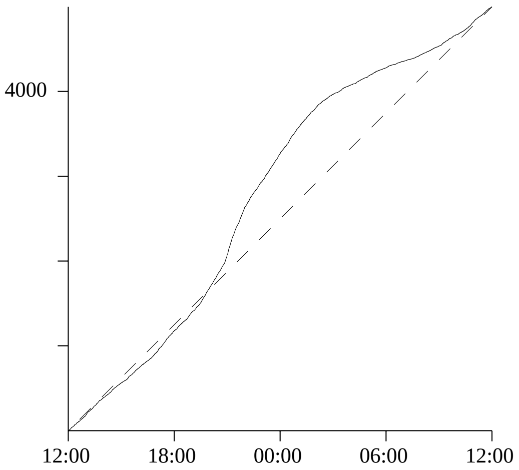

Fig 6:

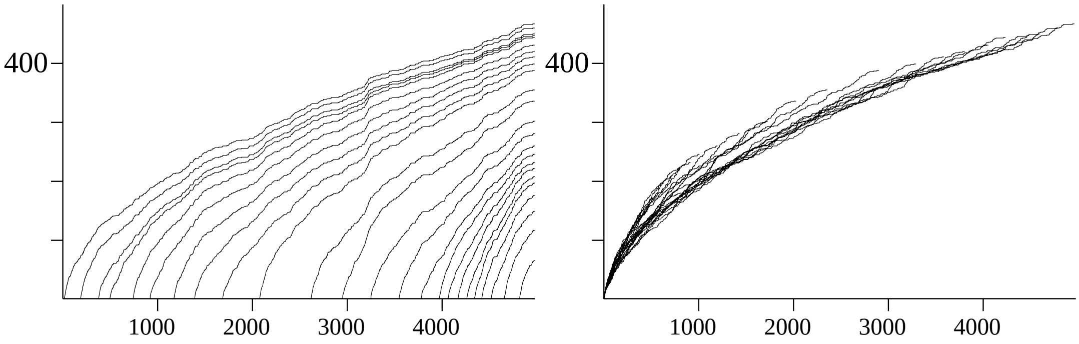

Fig 7:

Using (41) in (26), with the Pareto distribution (32) for λ,

$$x_C(t) = N y_C(t) + 1 \simeq N - N \int_{\mathbb{R}_+} e^{-w\, S^{(N)}(t)/Z(N)}\, \lambda(dw) =$$

$$N - N e^{-S^{(N)}(t)/(N^{1/b}\zeta_N(1/b))} + (\frac{S^{(N)}(t)}{\zeta_N(1/b)})^b\, \Gamma(1-b, \frac{S^{(N)}(t)}{N^{1/b}\zeta_N(1/b)}) =: x_b^{(N)}(S^{(N)}(t)), \tag{43}$$

where $\zeta_N(z) = \sum_{i=1}^{N} \frac{1}{i^z}$. Denote the data of size $n_d = 70140$ given in Fig. 7 by (s_i, x_i), $i = 1, 2, \cdots, n_d$. We performed a statistical fit of the data to (43), by minimizing $E = \sum_{i=1}^{n_d} \frac{(x_i - x_b^{(N)}(s_i))^2}{x_b^{(N)}(s_i)}$, with $N = 697$, and obtained $b = 0.872 \pm 0.002$ (90% CL). Apparently, we have a good single parameter fit to the data, which suggests that the practical assumption (24) is good.

We note that a smaller value of b was obtained for 2ch.net in [14] (with a different set of data). The data used in [14] was small in size, because the data was collected manually in those times, and also, to avoid influence of day-night activity difference, the data was for a short time period, hence the result in [14] is less reliable compared to the present result.

We also note that we have $b < 1$, consistently with observation for Amazon.co.jp,

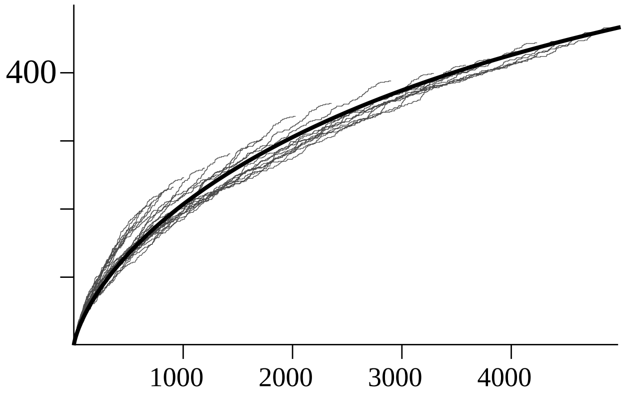

Fig 8:

where we obtained $b = 0.809$. This shows that, as in Amazon.co.jp, the popularity of subjects is concentrated on a small number of threads in 2ch.net.

References

[1] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More,* Hyperion Books, 2006.

[2] H. Bauer, *Measure and Integration Theory,* De Gruyter, 2001.

[3] P. Billingsley, *Convergence of probability measures*, John Wiley and Sons, 2nd ed., 1999

[4] J. R. Bitner, *Heuristics that dynamically organize data structures,* SIAM J. Computing **8** (1979) 82–110.

[5] G. Blom, L. Holst, *Embedding procedures for discrete problems in probability,* Math. Scientist **16** (1991) 29–40.

[6] P. J. Burville, J. F. C. Kingman, *On a model for storage and search,* J. Appl. Probab. **10** (1973) 697–701.

[7] F. R. K. Chung, D. J. Hajela, P. D. Seymour, *Self-organizing sequential search and Hilbert's inequalities,* J. Computer and System Sciences, **36** (1988) 148–157.

[8] R. Fagin, *Asymptotic miss ratios over independent references,* J. Computer and System Sciences, **14** (1977) 222–250.

[9] J. A. Fill, *An exact formula for the move-to-front rule for self-organizing lists,*

J. Theor. Probab. **9** (1996) 113–160.

[10] J. A. Fill, *Limits and rates of convergence for the distribution of search cost under the move-to-front rule,* Theoretical Computer Science, **164** (1996) 185–206.

[11] J. A. Fill, L. Holst, *On the distribution of search cost for the move-to-front rule,* Random Structures and Algorithms **8** (1996) 179–186.

[12] Y. Hariya, K. Hattori, T. Hattori, Y. Nagahata, Y. Takeshima, T. Kobayashi, *Stochastic ranking process with time dependent intensities,* Tohoku Mathematical Journal (2011), to appear.

[13] K. Hattori, T. Hattori, *Existence of an infinite particle limit of stochastic ranking process,* Stochastic Processes and their Applications **119** (2009) 966–979.

[14] K. Hattori, T. Hattori, *Equation of motion for incompressible mixed fluid driven by evaporation and its application to online rankings,* Funkcialaj Ekvacioj **52** (2009) 301–319.

[15] K. Hattori, T. Hattori, *Mathematical analysis of long tail business using stochastic ranking processes,* preprint, 2008.

[16] K. Hattori, T. Hattori, *Sales ranks, Burgers-like equations, and least-recently-used caching,* Kokyuroku Bessatsu (2011), to appear.

[17] W. J. Hendricks, *The stationary distribution of an interesting Markov chains,* J. Appl. Probab. **9** (1972) 231–233.

[18] P. R. Jelenković, *Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities,*

Ann. Appl. Probab. **9** (1999)430–464.

[19] P. R. Jelenković, A. Radovanović, *Least-Recently-Used caching with Dependent Requests,* Theoretical Computer Science **326** (2004) 293–327.

[20] P. R. Jelenković, A. Radovanović, *The Persistent-Access-Caching algorithm,* Random Structures and Algorithms, **33-2** (2008) 219–251.

[21] J. F. C. Kingman, *Random discrete distributions,* J. Roy. Stat. Soc. Ser. B **37** (1975) 1–22.

[22] G. Letac, *Transience and recurrence of an interesting Markov chain,* J. Appl. Probab. **11** (1974) 818–824.

[23] J. McCabe, *On serial files with relocatable records,* Oper. Res. **13** (1965) 609–618.

[24] R. Rivest, *On self-organizing sequential search heuristics,* Comm. ACM **19** (1976) 63–67.

[25] E. R. Rodrigues, *Convergence to stationary state for a Markov move-to-front scheme,* J. Appl. Probab. **32** (1976) 768–776.

[26] M. L. Tsetlin, *Finite automata and models of simple forms of behaviour,* Russian Math. Surv. **18** (1963) 1–27.